DIAN: A Novel Algorithm for Genome Ontological Classification

Yannick Pouliot, Jing Gao, Qiaojuan Jane Su, Guozhen Gordon Liu, and Xuefeng Bruce Ling[1,2]

DoubleTwist, Inc., Oakland, California 94612, USA

Faced with the determination of many completely sequenced genomes, computational biology is now faced with the challenge of interpreting the significance of these data sets. A multiplicity of data-related problems impedes this goal: Biological annotations associated with raw data are often not normalized, and the data themselves are often poorly interrelated and their interpretation unclear. All of these problems make interpretation of genomic databases increasingly difficult. With the current explosion of sequences now available from the human genome as well as from model organisms, the importance of sorting this vast amount of conceptually unstructured source data into a limited universe of genes, proteins, functions, structures, and pathways has become a bottleneck for the field. To address this problem, we have developed a method of interrelating data sources by applying a novel method of associating biological objects to ontologies. We have developed an intelligent knowledge-based algorithm, DIAN, to support biological knowledge mapping, and, in particular, to facilitate the interpretation of genomic data. In this respect, the method makes it possible to inventory genomes by collapsing multiple types of annotations and normalizing them to various ontologies. By relying on a conceptual view of the genome, researchers can now easily navigate the human genome in a biologically intuitive, scientifically accurate manner.

Biologists have never before been exposed to such vast amounts of sequence data as that from the human genome and a variety of model organisms. This development now raises the issue of how to interpret the meaning of the genome on the basis of prior biological understandings. Annotation tasks, such as the prediction of protein function and structure, are essential to this process and are by no means completely robust. Furthermore, the integration of historical domain knowledge accumulated in individual research fields with these sequence and structural annotations is becoming increasingly complex and difficult. The size, diversity, and complexity of the data, which include biological sequence information itself, third party or in-house annotation, and information from the scientific literature, are responsible for these difficulties. Another reason relates to the lack of data and information normalization, because the data repositories are often poorly designed, particularly in the case of older repositories. Furthermore, data processing procedures vary substantially, and the underlying semantic and data models are moving targets. Finally, there is the extreme specialization of research fields.

Despite these problems, model organism studies and associated DNA and protein sequence data sets have revealed a high degree of sequence and functional conservation between organisms (Chervitz et al. 1999). Similarly, the accumulated protein structure data have shown that the number of protein folds is probably limited (Bowie et al. 1991). The limited number of biological roles, protein functions, and structural types enable a common language for annotation, which is beginning to be implemented by biocomputational ontology engineering (Riley 1993; Baker et al. 1999; Ashburner et al. 2000; Karp 2000). Ontologies provide an ideal mechanism of organization of biological data at the conceptual level by providing a framework for data whose properties are otherwise nonnormalized. "Normalization" is used here to refer to a state in which several types of signifiers ultimately express the same concept, and in which a concept is defined as a generic abstraction derived from instances. An example of a concept is the notion of "cell adhesion molecules", to which specific types (instances) of proteins such as cadherins, neural cell adhesion molecules, and integrins are conceptually associated. The proper assignment of DNA and protein sequences to ontologies therefore leverages the rigor of the underlying concepts networked within these ontologies, and enables computations that would otherwise be unreliable due to the variability of terms used to described biological data in most of the biocomputational databases. For example, ontology-based querying can enable the retrieval of DNA and protein sequences based on biological concepts rather than relying on keyword or synonym searches, which are inherently unreliable due to their present nonnormalized nature, therefore greatly hampering effective computing (Attwood 2000).

Here we describe DIAN, an ontology assignment algorithm that assigns concepts to source records or, more generally, to biological objects within a database, and supports their querying using concepts rather than keywords. The algorithm supports a variety of ontologies for biological role, protein function, and protein structure, whereby each ontology is implemented on a knowledge base established via computer-assisted human curation of the protein universe. DIAN has the necessary throughput capacity to annotate entire genomes, transcriptomes, and proteomes onto any number of ontologies. The DIAN algorithm, together with the precom-

[1]Present address: Tularik, Inc., 2 Corporate Drive, South San Francisco, CA 94080, USA.
[2]Corresponding author.
E-MAIL XLING@tularik.com; FAX (650) 825-7400.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.183301.

11:1766–1779 ©2001 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/01 $5.00; www.genome.org

puted DIAN annotation database and its associated utilities, enables users to retrieve, summarize, and predict the higher order properties of biological objects, therefore increasing their information content. Overall, DIAN is intended to facilitate the navigation of genomic data repositories in a biologically intuitive, scientifically accurate manner.

RESULTS AND DISCUSSION

Biologists rely heavily on databases and search tools such as the National Center for Biotechnology's Entrez system to search and identify records containing information associated with biological objects such as protein structures and biological sequences (Wheeler et al. 2001). However, when computing on such information, most query systems suffer from the limitations inherent to the annotations associated with these objects. Even in highly curated databases such as the SWISS-PROT database of protein information (Bairoch 1991), there remains significant variability in the descriptors present in these source records. This is because there are many legitimate ways of describing biological concepts. Furthermore, even when the data are curated by experts, a variety of factors introduce variability in the quality and comprehensiveness of these annotations. Thus, when querying annotation databases, conventional search tools encounter fundamental limitations, such that they cannot return records in a reliable manner unless a complete set of descriptors known to be present in the targeted records is provided in the query. This, of course, rarely is the case.

DIAN is designed to enable the querying of popular biological databases in such a way that the limitations associated with the original source records of these databases can be partially overcome. This is accomplished by having the operator query biological ontologies for records associated with these ontologies, rather than querying the source records directly (for details, see supplementary material at http://www.genome.org). The primary algorithm used by DIAN for associating records to ontologies relies on a domain-based approach that does not depend on the presence of annotations in the source record, thus bypassing the limitations associated with these annotations. In addition, because of this approach, DIAN often makes suggestive assignments, whereby proteins are predicted to belong to ontological nodes in the absence of definitive information.

For these reasons, when performed using conventional keyword-based search engines, the queries described in Table 1 will fail to return a fraction of records because of an absence of matching annotations or because of the indirectness of these annotations (i.e., hyperlinked records). Three such cases of records that would otherwise not have been returned without DIAN are illustrated in Table 1. They involve two novel genes, one with predicted functional information listed in the source record and one without such information, as well as one well-characterized gene. In case 1, DIAN identified a gene with no known functional activity by predicting the cellular role and protein function of a sequence on the basis of its pattern of protein domains. UniGene was queried for records involved in the apoptotic Cellular Role. DIAN returned a record from the UniGene database where no functional information is available regarding this sequence, such that this record would not have been identified by keyword-based querying (Table 1). It is only after consulting the SWISS-PROT record linked to this UniGene entry that an apoptotic function is uncovered. Case 2 concerns the prediction of a cellular role for a hypothetical gene in SWISS-PROT in which putative functional information is available (zinc finger; DNA binding) but where the annotation does not specify a cellular role. In this case, DIAN predicted an involvement in the "RNA synthesis/transcription factor" Cellular Role node. In case 3, DIAN predicted a novel property for a highly characterized gene. Here, UniGene was queried for records involved in the apoptotic Cellular Role. The gene coding for the protein associated with the Wiskott-Aldrich syndrome (WAS; Derry et al. 1994) was one of the hits returned by this query. The WAS protein is thought to be involved in signal transduction, yet there is no indication of an apoptotic role in any of the records associated with this gene, including the associated SWISS-PROT and OMIM records. However, indications suggestive of a possible apoptotic role were found in these sources. On subsequent analysis of the scientific literature associated with WAS and its *Drosophila* ortholog, several publications were uncovered that strongly substantiate a recently discovered apoptotic role for this gene (Rawlings et al. 1999; Rengan et al. 2000; Ben-Yaacov et al. 2001). Beyond the performance of DIAN in returning records otherwise unretrievable, the combination of ontology-based and Boolean operators (e.g., NOT, AND, OR) enables users to query databases in a biologically meaningful manner rather than to submit to unfamiliar querying syntaxes and the vagaries of unstructured data. For example, using DIAN it is possible to formulate directly the following questions in a simple manner: Are there cytokines involved in the apoptosis biological process? Are there proteins harboring the caspase domain that are involved in apoptosis? What receptors are associated with apoptosis? What proteins are both apoptosis-associated and DNA-associated in terms of cellular role? (i.e., proteins that might perform an apoptotic role via DNA binding). Such questions cannot be addressed if the contents of annotation databases have not been normalized to various biological concepts, and furthermore, comprehensive biological query cannot be performed reliably when accomplished exclusively by using a simple keyword-search approach, as seen in most public databases.

Organization of Biological Data Using Ontologies

An ontology is a specification of a conceptualization that provides a written, formal description of a set concepts and their relationships within a domain of interest (Karp 2000). Ontologies are object-oriented data structures that use object composition and inheritance as techniques to encapsulate conceptual relationships.

In biology, there are two kinds of relationships between conceptual objects to be represented: inheritance and compositional relationships. Inheritance hierarchies model IS-A relationships among base and derived conceptual objects. This is because a derived object IS-A type of base object. In contrast, composite objects, that is, objects that contain other objects as members, model HAS-A relationships. This is because the container object HAS-Another as its member component. For example, in the Gene Ontology (GO) ontologies (Ashburner et al. 2000), the Cellular Component ontology relies on HAS-A compositional relationships, whereas the Molecular Function ontology uses IS-A inheritance relationships. In this way, the granularity and richness of the universe of biological concepts can be modeled by ontologies.

To encapsulate biological conceptual objects and support the goal of concept-based searching, the DIAN algorithm segments the spaces of protein function, biological role, and pro-

Table 1. Example Queries that Cannot be Resolved Accurately by Conventional Querying Systems

Case type	Identifier / Annotation	Source record: DIAN system annotation				
		EGAD celluar role	Protein function	Enzyme classification	DoubleTwist biological role	Structure (SCOP)
Case 1: Novel gene with no predicted function	UniGene Hs. 104305 / None	•cell/organism defense •homeostasis •apoptosis •cell division •apoptosis	None	None	•Non-immune cell defense •Apoptosis	•All alpha proteins •DEATH domain •DEATH domain •DEATH domain
Case 2: Hypothetical gene with predicted function	SWISS-PROT P39959 / Putative Zinc Protein	•gene/protein expression •RNA synthesis •transcription factors	•DNA or RNA associated proteins	None	•Genome structure and Gene expression •Transcription factors	•Small proteins •Classic zinc finger, C2H2 •Classic zinc finger, C2H2 •Classic zinc finger, C2H2
Case 3: Known gene with novel predicted function	UniGene Hs. 2157 / Wiskott-Aldrich syndrome protein / OMIM: 30100 SWISS-PROT: P42768	•cell division •apoptosis	•Enzymes •Transferase •Post-translational modifications	•Transferases	•Non-immune cell defense •Apoptosis	•All beta proteins •PH domain-like •PH domain-like •Enabled/VASP homology 1 domain (EVH1 domain)

Three illustrative cases of records that cannot be returned by conventional keyword-based querying systems but that were returned by DIAN are described here.

tein structure using a collection of ontologies. Although HAS-A relationships should be supportable, in this study we rely exclusively on IS-A ontologies as a paradigm to show the DIAN methodology. Computationally, ontologies take the form of a graph, a tree being a special form of a graph. A node always inherits the properties of all parental nodes, such that a complete description of the biochemical function of a protein involves starting the path from the leaf to the root of the tree. The first three levels of the PROSITE Protein Function ontology are used to illustrate this conceptualization (Fig. 1). Starting from the root of the tree (level 0), each level describes biochemical protein function in increasingly greater detail. In this illustration, six proteins were assigned to the transferase node. Because this is a protein function ontology, proteins can belong to different families and species and yet be assigned to the same node. By providing standard classification data structures, ontologies are ideal in providing a common platform for annotation and therefore promoting reuse across different informatics systems and research fields. Because the focus of this paper is on a methodology for assigning protein sequences to ontologies, the relative merits of individual ontologies are addressed only briefly.

Choice of Ontologies

Because an ontology is essentially a specification of conceptualization (Karp 2000), the choice and quality of the chosen ontologies are essential in ensuring the integrity of encapsulating the biological data. To support the conceptualization of protein functions, biological roles, and cellular processes, substantial attention has been devoted, both in academia and industry alike, to the development of various ontologies to meet these needs. Examples include the enzyme commission classification system (Commission on Biochemical Nomenclature and International Union of Biochemistry. Standing Committee on Enzymes 1973; International Union of Biochemistry and Molecular Biology. Nomenclature Committee and Webb 1992; International Union of Biochemistry. Nomenclature Committee and Commission on Biochemical Nomenclature 1979; International Union of Biochemistry. Nomenclature Committee et al. 1979; International Union of Biochemistry. Nomenclature Committee et al. 1984; International Union of Biochemistry. Standing Committee on Enzymes 1965); the *Escherichia coli* Protein Function ontology (Riley 1993); the EcoCyc system for *E. coli* metabolic pathway (Karp 2000); the PROSITE ontology of domain biological functions (Hofmann et al. 1999); the GO ontologies (Ashburner et al. 2000); the KEGG system for the classification of genes according to pathway information (Ogata et al. 1999); RIBOWEB (Chen et al. 1997); and the TIGR expressed gene anatomy database (EGAD, http://www.tigr.org/tdb/egad/egad.shtml). Similarly, to facilitate the understanding and access to information of protein structures, several protein structure classifications have been constructed (Murzin et al. 1995; Orengo et al. 1997). Despite these efforts, there is still no accepted ontology with the necessary robustness, comprehensiveness, and level of detail to satisfy the demands of genome annotation, although this is an implied goal of the GO project.

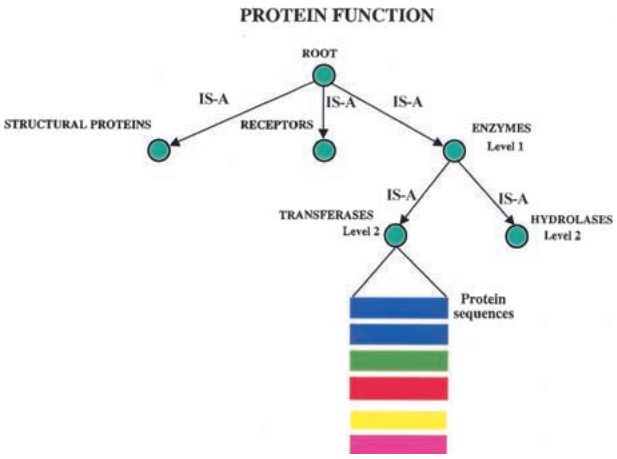

Figure 1 Defining ontologies. Ontologies represent a specification of a domain of knowledge expressed in the structure of mathematical graphs (a tree being a special form of a graph). Connecting lines represent the relationship between the nodes, specifically IS-A relationships. Known protein functions are assigned to nodes (represented by circles) within the ontology graph. A child node always inherits the properties of all parent nodes, such that a complete description of the biochemical function of a protein involves retracing the path from the leaf to the root of the tree.

Given these limitations, and in the absence of the GO ontologies, we originally chose to rely on various publicly available ontologies, in addition to deriving the DoubleTwist Biological Role Ontology (Table 1). For Protein Function, DIAN supports the PROSITE Protein Function and the enzyme commission classification. Given its rigor, comprehensiveness, and rapid evolvement, the GO ontologies, including its three components (molecular function, biological process, cellular components), are expected to be integrated within DIAN in the foreseeable future. For the Cellular Role of proteins, the TIGR expressed gene anatomy database (EGAD) ontology and DoubleTwist Biological Role ontology are supported. Although not very comprehensive, the EGAD ontology currently is the only publicly available ontology designed to inventory human expressed genes. The DoubleTwist Biological Role ontology was derived from Riley's Protein Function ontology (Riley 1993) and has been designed for the concise conceptual encapsulation of the biological role of the human gene to enable comprehensive human genome assignment. As for protein structure classification, the Structure Classification of Proteins (SCOP) ontology was selected because SCOP is sequence-based and its classifications provide a detailed and comprehensive description of the structural and evolutionary relationships of proteins of known structure (Murzin et al. 1995).

Architectural Design

DIAN (Fig. 2) integrates several databases through algorithms that perform the ontological assignment of proteins on the basis of two distinct principles. The first algorithm (vocabulary-based mapping) relies on the recognition of vocabulary within a source record from a database of protein annotations. The second algorithm (domain-based mapping) assigns protein sequences based on the detection of protein domains and does not rely on preexisting sequence annotation.

DIAN has several subcomponents in support of these functions: a knowledge base of assignments of SWISS-PROT proteins to ontologies; two databases that provide operational definitions for each ontological node, based either on vocabulary or the assignment of protein domains; two assignment algorithms for assigning proteins on the basis of either vocabulary or the presence of protein domains; and lastly a data

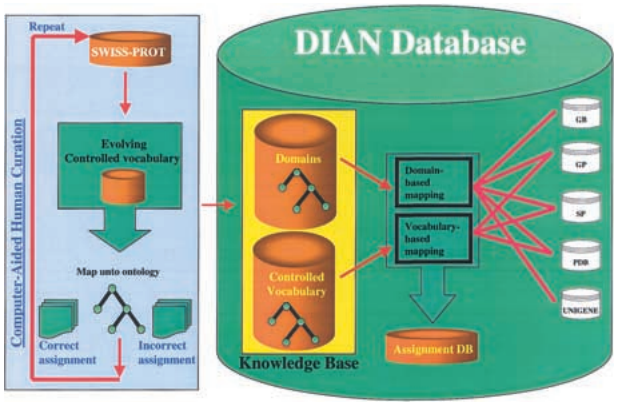

Figure 2 DIAN overview. (*Left*) computer-aided human curation process for the assignment of SWISS-PROT sequences to the Protein Function and Cellular Role ontologies; (*right*) application of ontologies to organize biological annotation databases. Multiple ontologies, each representing a body of biological knowledge, are stored in the DIAN database. Individual source records stored in biological sequence and structure annotation databases are associated with one or more ontologies via domain-based and/or vocabulary-based mapping, such that they can be queried simultaneously across multiple ontologies. (GB) GenBank; (GP) GenPept; (SP) SWISS-PROT; (DB) database.

indexing and retrieval engine to support user queries. Each subcomponent is described in the following sections.

Development of the DIAN Knowledge Base

An essential component in DIAN is a knowledge base derived from a computer-aided human curation process that associates entries of the SWISS-PROT database to ontologies. SWISS-PROT is known for its high-quality curated annotations of protein sequences and minimal level of redundancy (Bairoch and Apweiler 2000). Although most sequence databases provide SWISS-PROT links to leverage its high-quality annotations, accurate and comprehensive classification of SWISS-PROT entries onto Protein Function and Cellular Role ontologies has not been achieved. DIAN relies on this knowledge base as a foundation to define parameters and data sets to support the computational assignment of proteins to ontologies.

During the early phase of the development of this knowledge base, we attempted to rely on preexisting links between SWISS-PROT and other publicly available databases to determine whether these links could be used directly to associate SWISS-PROT records to ontologies. It was found that this superficially straightforward method of assignment is error prone, and that the resulting coverage of SWISS-PROT was not comprehensive. Instead, the assignment of SWISS-PROT to the Protein Function and Cellular Role ontologies stored in the knowledge base was achieved through a computer-aided manual curation process (illustrated in Fig. 2). A group of scientific curators was assembled to manually assign SWISS-PROT sequences to the DIAN Protein Function and Cellular Role ontologies by matching the functional annotation of each SWISS-PROT record to the definition of each node in a given ontology. To ensure the high accuracy of this underlying data set, we analyzed only the subset of SWISS-PROT proteins that are full length and have been characterized biochemically. This resulted in the initial assignment of over 40,000 proteins to the DIAN ontologies. Subsequent to this manual curation process, a database of controlled vocabulary was evolved from the assignment,in which for each ontological node, essential keywords were extracted from the annotations of the SWISS-PROT proteins assigned to the node. To enhance the selectivity and sensitivity of each definition, this data set was used to partition SWISS-PROT according to records that are either positively or negatively assigned to a node. Each set of partitioned SWISS-PROT records was examined thoroughly by curators to identify false positive records in the positive pool, and records characterized as false negatives in the negative pool. This information was then used for a second round of keyword refinement as feedback data in generating a subsequent, more refined set of controlled vocabulary. This process was repeated until no further additional identifiable false positives could be detected. Once this data set stabilized for all nodes, the SWISS-PROT-Ontology assignment table was finalized, resulting in the assignment of over 84% of SWISS-PROT (for further details, see on-line supplementary Table 2B at http://www.genome.org). This information was added to the knowledge base, such that it now provides an operational definition that expresses the knowledge associated with each node. The knowledge base was later used as the foundation for the development of nodal signatures, and along with periodic verifications of the selectivity and sensitivity on new releases of SWISS-PROT, it ensures the continued assignment of SWISS-PROT entries to the Protein Function and Cellular Role ontologies as SWISS-PROT evolves.

Development of Nodal Signatures

To classify biological sequence annotations by assigning them to ontologies, we developed annotation signatures for each node of the supported ontologies. Such nodal signatures provide the operational definitions used by the DIAN assignment algorithms to recognize properties in protein sequences, such that sequences from input databases can be assigned to ontologies. Two kinds of nodal signatures are used in the DIAN algorithm: signatures based on either controlled vocabulary or protein domain profiles. A protein domain is here defined as an independent structural unit, which can be found alone, or in conjunction with other domains. Domains are often the mediators of the biochemical functions of proteins, although a substantial fraction of domains appears to play structural roles only. For this and other reasons, not all domains can be used as nodal signatures. For the Protein Function and Cellular Role ontologies, controlled vocabulary databases are used to efficiently collapse protein annotations present in source records and to assign these records to ontologies, as was done when assigning SWISS-PROT sequences to ontologies during the development of the knowledge base. This controlled vocabulary is expected to accurately classify sequence via annotations preexisting in the source records as long as the quality of these annotations is comparable to that of SWISS-PROT. Although sensitive enough to capture input sequence annotations under most circumstances, this approach is essentially a keyword-matching mechanism that may incorrectly assign records to ontologies as compared with the actual sequence annotation. This is an expected consequence of the process by which nodal vocabularies are derived. For example, it is possible for both a kinase substrate and a kinase enzyme to become assigned to the same ontology kinase node, when in fact only kinase enzymes should be assigned to this node. This is a consequence of the difficulty of defining assignments on the basis of vocabularies alone.

Of larger consequence is the intrinsic quality of the annotations associated with a sequence to be assigned, because annotations in most sequencing projects are transferences obtained through sequence similarity alignments with characterized gene or proteins. This can lead to so-called "multiple-linkage" errors during the annotation transfer process, which creates misleading annotations due to the localization of the alignment in a region with low functional information content (e.g., a region devoid of a functional domain). Therefore, an additional assignment algorithm was derived to compensate for this well-known problem by relying only on the presence of domains within protein sequences or the translation of DNA sequences into proteins. Whereas evolutionarily and functionally related protein sequences can diverge significantly through evolution, three-dimensional substructures, such as motifs, domains, and active sites, can remain largely unchanged (Gusfield 1997). As a result, protein domain profiles compiled from multiple sequence alignments can enable more accurate representation of protein families and superfamilies. Furthermore, such conserved sequence features are highly correlated with structure and function. As a result of the success of the protein profiling methodology, several protein domain and motif databases have been built: PFAM (Sonnhammer et al. 1998), PROSITE (Bairoch 1991), PRODOM (Corpet et al. 1998), DOMO (Gracy and Argos 1998), EMOTIF/EMATRIX (Nevill-Manning et al. 1998; Wu et al. 2000), BLOCKS (Henikoff et al. 1999), PRINTS (Attwood et al. 1997). Although in the current DIAN algorithm we have chosen to rely on domains provided by the PFAM database because of its extensive coverage and the richness of its associated annotations, other domain or motif databases can be integrated in the same fashion.

Because of the close relationship between a given protein domain and the function and structure of a protein that harbors this domain, the ontological classification of protein sequences using well-chosen protein domains can be achieved by using an effective balance between the specificity and sensitivity of individual domains. A filtering algorithm was therefore developed to select domains qualified to function as nodal signatures to be used in assigning proteins to ontologies. Comprehensive analyses of the DIAN knowledge base for patterns of association between PFAM domains and SWISS-PROT sequences assigned to ontological nodes revealed frequent many-to-many relationships between domains and nodes. To promote specificity, it was therefore necessary to analyze all preliminarily assigned protein domains for the possibility of conversion to nodal signatures for a particular node. This was accomplished in the following way: For each of the protein domains in the source pool, the annotations of all SWISS-PROT sequences containing a particular protein domain were compared against the assignment of this sequence to a node, as maintained in the DIAN knowledge base. Second, if a set of annotations associated with sequences containing a given protein domain was found to be correlated with the description of the node, this domain was accepted as the annotation signature for that ontology node, as this domain is relatively specifically correlated with that node.

These concepts are illustrated in Figure 3 in the case of the Protein Function ontology. This ontology is expressed as a tree in which each node represents a concept and is associated with other concepts via an "IS-A" relationship. The root of this tree (level 0) is a generic function. Child nodes inherit the properties of their parent and express increasingly specific protein functions. For example, among the children of the root lies the "enzyme" node, which is defined as " biomolecules that can catalyze reactions." Associated with this node are keywords positively correlated with this function, such as "Oxidoreductase OR Transferase OR Hydrolase OR Lyase OR Isomerase OR Ligase". As a first step in the derivation of the DIAN knowledge base, proteins described in the SWISS-PROT database were assigned to the most specific nodes possible. Here, six proteins were assigned to the transferase node (Fig. 3). Two proteins belong to the same gene family and are of human origin, whereas all other proteins are from different gene families from various species. Various protein domains are present within these proteins, sometimes more than once in a given protein. Thus, a total of five distinct types of protein domains are present within the group of proteins assigned to the transferase node. However, only three types of domains are retained by DIAN as protein annotation signatures, because according to the DIAN knowledge base these domains are the only domains to be specifically associated with transferase-related functions. Thus, the two remaining domain types were rejected as annotation signatures because they are either not encoding a function related to transferases, or are purely structural domains not directly involved in protein function. In this way, any database of protein motifs or domains can in principle be integrated in the DIAN algorithm to derive ontological node signatures. The current DIAN implementation relies on protein domains from the PFAM database as its source of protein domains to be converted into ontological node signatures.

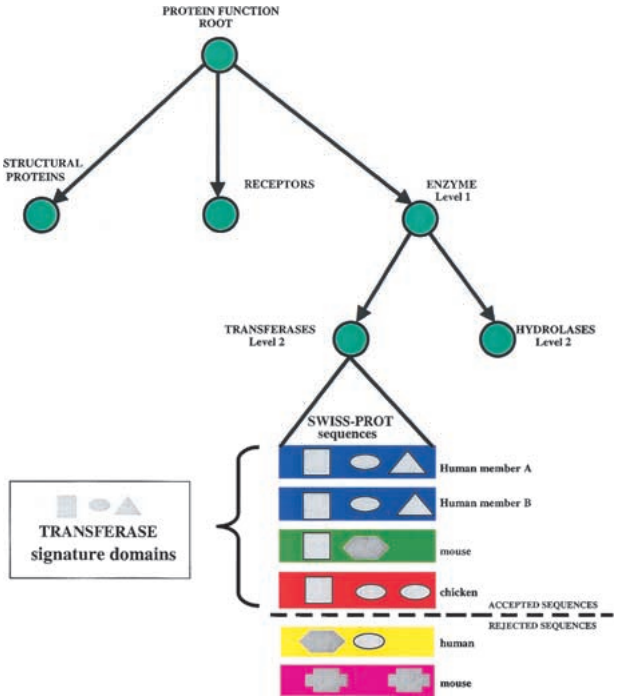

Figure 3 Converting protein domains into ontological nodal signatures. The Protein Function ontology is used here to illustrate the derivation and assignment of nodal annotation signatures. Proteins are depicted as rectangles; identical colors indicate membership to the same protein family in a given species, whereas the various protein domains are represented as geometrical shapes.

Based on overlaps between the annotations present in the 86,593 sequences of release 39 of SWISS-PROT and the

concepts associated with our ontological nodes, computer-aided human curation associated 73% of SWISS-PROT sequences to the PROSITE Protein Function ontology, 68% to the EGAD Cellular Role ontology, and 68% to the DoubleTwist Biological Role ontology. Overall, 205,694 keyword-based patterns and 1699 PFAM domains were compiled to represent the biological concepts associated with each ontology node.

Nodal signatures for the structural ontology were derived differently from the process described in Figure 3. This was achieved by profiling the SCOP domain sequences compiled by the SCOP consortium (Brenner et al. 1998), using selected protein domains from the PFAM database. Because high sequence similarity usually implies significant functional and structural similarity (Gusfield 1997), 824 PFAM domains were identified that are referenced in sequences of the SCOP domain database (S. Brenner, pers. comm.). These PFAM domains show strong sequence similarity to SCOP domains and were selected because they are likely to represent a similar structure in three-dimensional space.

Ontological Assignment Process

Two assignment algorithms are used to assign proteins to DIAN ontologies. This is achieved on the basis of either the presence of protein domains or the recognition of vocabularies within the source record. As shown in Figure 2, annotations in various biological sequence databases, including GenBank, SWISS-PROT, GenPept, PDB, and UniGene, are collapsed through either the domain-based or vocabulary-based algorithms into a centralized DIAN database. In cases where DNA sequences are operated on by the domain-based algorithm, a translation algorithm is applied, as DIAN only operates ultimately on protein sequences. Genomic DNA sequences are treated differently in this process because these sequences show very different properties from cDNA and proteins. In particular, sequence length can easily exceed a million characters. For this reason, it would therefore be incorrect to apply ontologies directly at the level of an entire genomic sequence. Thus, location coordinates are essential to segment genomic sequences into biologically meaningful ranges ("units") before further processing. If available in the source record, information specifying the presence of genes, derived ab initio or experimentally, are used to define the unit. However, in sequences derived from high-throughput sequencing projects (e.g., sequences from the GenBank HTG division), this information is frequently unavailable. In such cases, DIAN can use available gene predictions from algorithms such as GENSCAN (Burge and Karlin 1997) or GENEWISE (Birney and Durbin 2000) to locate the genes in the genomic sequence.

As mentioned earlier, another assignment approach applied by DIAN is based on the scanning of annotations associated with the input biological sequence using a vocabulary-based mapping process. This is accomplished by the application of a collection of keywords that serve as the ontology node annotation signatures, enabling the collapse of preexisting annotations and their assignment to ontological nodes. The input sequence annotations can be derived from sequence similarity information, domain profiling information, human curation, computation-derived annotations, third-party annotations, and so forth. Together, the domain-based and vocabulary-based algorithms are used by DIAN to annotate and classify sequences from input biological databases in a high-throughput manner.

DIAN Algorithm Evaluation

The sensitivity and selectivity of the DIAN algorithm were evaluated. Based on sequence similarity results, the vocabulary-based algorithm implicitly transfers existing annotations and assigns proteins to ontological nodes. However, this process suffers from two intrinsic types of errors: Because of the variability of vocabularies in the annotations, it is very difficult to identify and compensate for incorrect annotations during this annotation transfer process. Furthermore, multiple linkage errors are generated when annotations are wrongly transferred when the sequence similarity between both sequences is only present within core structural regions with low information content, rather than encompassing functional domains. However, the domain-based assignment algorithm is not susceptible to these problems. Thus, despite the observation that the domain-based algorithm generated less coverage than the vocabulary-based algorithm, the domain-based algorithm can make annotation assignments in the absence of preexisting annotations in the source records.

The accuracy of an ontological mapping algorithm such as DIAN is defined as the fraction of correct assignments made to the nodes of an ontology, both in terms of type I variations (assignments that should not have been made but are present) and type II variations (assignments that are missing and that should have been made). Here we use the terms types I and II "variation", rather than "type I error" and "type II error", to emphasize that providing exact error rates in this context is fundamentally impossible (see the following discussion of error measurements in this context). The accuracy of the DIAN algorithm was evaluated using three complementary approaches, summarized in Table 2. The construction of the underlying data sets is described in Figure 4. Detailed results of evaluations are documented in Table 3.

The DIAN assignments of well-characterized mouse sequences were compared with assignments made via an independent assignment process (method 2, Table 2). These assignments were provided by the Mouse Genome Database (MGD; Blake et al. 2000) using the Molecular Function and Biological Process ontologies from the Gene Ontology (GO) Consortium (Ashburner et al. 2000; http://www.geneontology.org). The application of GO ontologies to the mouse genome was chosen over that of other organisms such as *Drosophila* and others because of its closer relationship to human proteins and the bias in the SWISS-PROT database toward higher organisms. Because these ontologies are different from those currently supported by DIAN, a cross-referencing was first determined to enable comparisons of assignments. As shown in Figure 4B, comparing assignments made to ontologies is accomplished first by manually selecting nodes from a reference ontology for concepts that are shared between the ontologies. Because of the different levels of resolution supported by different ontologies, nodes at equivalent levels of resolution need to be identified. This results in some of the terminal nodes of one ontology being associated with middle nodes of the counterpart ontology. Furthermore, multiple nodes from one ontology may need to be selected to represent the concepts associated with a single node from the counterpart ontology (indicated by purple nodes from the reference ontology, all of which are conceptually equivalent to a single node from the DIAN ontology). Thus, the node associated with the INHIBITORS concept on level 3 of the DIAN ontology is conceptually equivalent to the APOPTOSIS INHIBITORS and ENZYME INHIBITORS nodes

Table 2. Methodologies Involved in the Evaluation of DIAN: Strengths and Weaknesses

Approach number	Approach type	Description	Strengths	Weaknesses
1	Manual verification of assignments made to selected proteins.	In-depth review by domain experts of assignments made to well-understood proteins.	Extensive human expertise can confirm assignments made by method and substantiate its effectiveness.	Suffers from lack of comprehensiveness; biased in favor of well-understood proteins.
2	Comparisons with other assignment data sets using a test set of sequences.	Evaluation of sequence assignments made to cross-referenced ontologies using different methods.	Presence of extensive shared assignments for numerous proteins lends credence to the method under evaluation.	Assumes that the reference ontology can be treated as a standard of comparison; in practice, this is not the case. Results in the identification of weaknesses in both the test and reference ontology. Manual review is required to evaluate unbalanced assignments.
3	Comparisons between orthologs.	Verification that assignments made to closely related orthologs are balanced, (i.e., nearly identical).	Strong expectation that balanced assignments will be made.	Although orthologs share functions, even orthologs from closely related species don't necessarily have identical functions, resulting in unbalanced assignments; manual review is required to evaluate unbalanced assignments.

and subnodes on levels 6 and 8 and lower of the reference ontology. Other problems arise from the differing extent of coverage between ontologies, which can obscure the interpretation of the comparison. In this example, there are several more proteins mapped to the DIAN ontology than to corresponding nodes of the reference ontology. Some proteins are mapped to both ontologies (green area, where individual pair members are indicated by double arrows), whereas other proteins are mapped only to a single ontology (red area). Within the latter, a manual review will find that some proteins are correctly mapped (blue rectangles), whereas others are incorrectly mapped (yellow rectangles). Lastly, there can be variations in the comprehensiveness of assignments made to individual proteins, such that only a fraction of the properties associated with a single protein are assigned to an ontology (data not shown). Detailed results of this evaluation are listed in Table 3A and 3B.

A number of intrinsic problems were identified from our

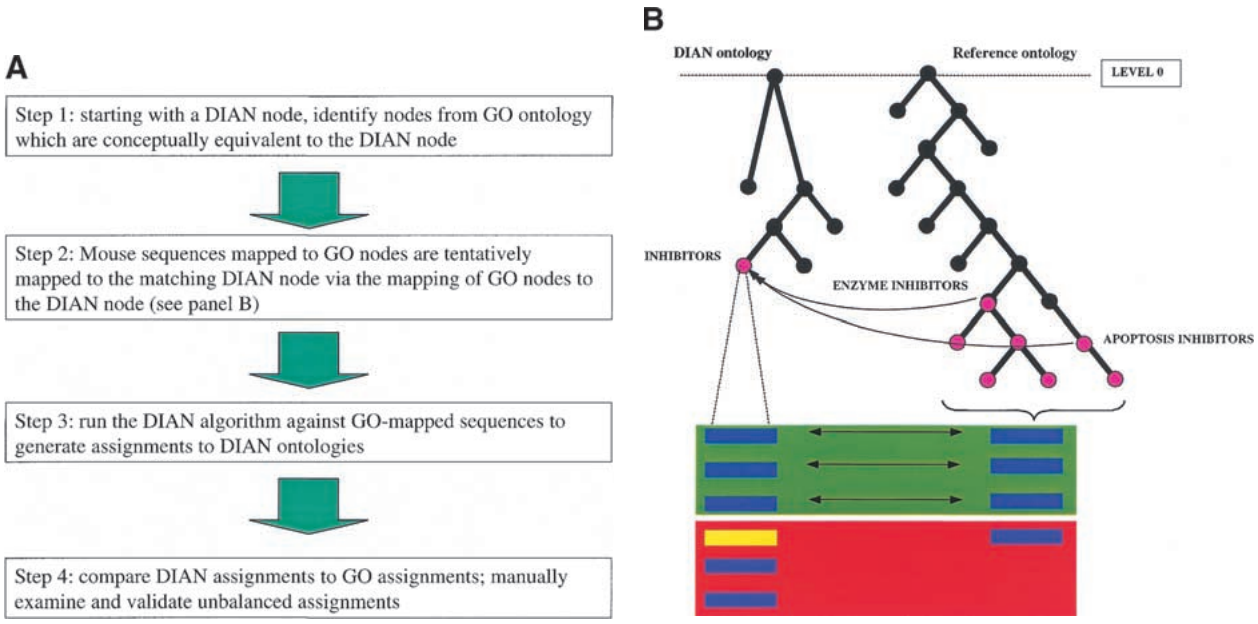

Figure 4 Validation approaches. (*A*) Evaluating the effectiveness of DIAN by comparing assignments made to a reference ontology. Selected nodes from Gene Ontology (GO) ontologies were manually associated with nodes in the DIAN ontologies. Sequences assigned to these GO nodes by the MGI were processed by the DIAN pipeline to compare the assignments made by DIAN with those made by MGI. (*B*) Associating nodes and sequences from a reference ontology to a DIAN ontology for comparative evaluation. To estimate the error rates associated with the DIAN assignment algorithms, we compared mouse sequences mapped via DIAN (*A*) with assignments made to GO ontologies by MGI.

Table 3. Comparison between DIAN and MGI Ontological Assignments

Results from the comparative approach are shown. A number of intrinsic problems were identified from this approach, such that type I and type II variances described here are for comparative purposes only and cannot be interpreted strictly as type I and II errors.

Table 3A. Comparing Assignments Made to the Cellular Role Ontology

Concept	DIAN node number	Highest level matching GO modes	Present in			Variation		Sensitivity	Selectivity
			DIAN and GO	DIAN only	GO only	Type I	Type II		
Chromosome structure	1.1	GO:0007001;GO:0006323	7	4	4	0.267	0.267	0.636	0.636
Transcription factors	1.4	GO:0003700	59	54	38	0.358	0.252	0.608	0.522
DNA duplication	3.2	GO:0006260;GO:0003964	10	2	2	0.143	0.143	0.833	0.833
Cell-cell adhesion	5.2	GO:0007155	35	15	14	0.234	0.219	0.714	0.700
Transcription factors	9.1.1.1	GO:0008135	1	0	1	0.000	0.500	0.500	1.000
Microtubule	6.2	GO:0007017	9	0	2	0.000	0.182	0.818	1.000
DNA repair	8.1	GO:0006281	14	2	9	0.080	0.360	0.609	0.875
Programmed cell death	8.2	GO:0006915	14	7	23	0.159	0.523	0.378	0.667
Channel and transporter	4.6	GO:0006810;GO:0005216	47	8	27	0.098	0.329	0.635	0.855
Amino acid metabolism	9.2	GO:0006519	4	1	7	0.083	0.407	0.476	0.625
Stress response	8.4	GO:0006950	5	6	55	0.091	0.833	0.083	0.455
Nucleotide metabolism	9.4	GO:0006140:0006205 GO:0006143	0	2	7	0.222	0.778	0.000	0.000
Cofactor metabolism	9.5	GO:0006731	4	0	3	0.000	0.429	0.571	1.000
Total DIAN and GO: 229 Total DIAN only: 113 Total GO only: 214 Total: 556 Average type I: 0.203 Average type II: 0.385 Sensitivity: 0.517 Selectivity: 0.670									

DIAN assignments made to a group of well-characterized, nonredundant mouse sequences were compared to assignments made by the MGI to the GO Process and Function ontologies. GO modes corresponding to DIAN nodes are listed, along with the abbreviated essential concept from the DIAN Role ontology. For brevity, only the highest level GO nodes are listed. The number of sequences whose assignment is shared to both sets of ontologies is indicated (DIAN and GO), as well as the number of sequence assignments which differed (DIAN only, GO only). These numbers are used to calculate Type I and II variation using the following equations: Type I variation = DIAN only/(DIAN and GO + DIAN only + GO only); Type II variation = GO only/(DIAN and GO + DIAN only + GO only); Sensitivity = DIAN and GO/(DIAN and GO + GO only); Selectivity = DIAN and GO/(DIAN and GO + DIAN only). Sensitivity is defined as the ability of the DIAN algorithm to make what are believed to be all possible correct assignments. Selectivity is defined as the ability of the DIAN algorithm to not make what is believed to be an incorrect assignment.

Table 3B. Comparing Assignments Made to the Protein Function Ontology

Concept	DIAN node number	Highest level matching GO modes	Present in			Variation		Sensitivity	Selectivity
			DIAN and GO	DIAN only	GO only	Type I	Type II		
Hormones and active peptides	10	GO:0005179;GO:0005103 GO:0005104;GO:0005105 GO:0005106;GO:0005109 GO:0005110;GO:0005111 GO:0005112;GO:0005113 GO:0005114;GO:0005115 GO:0005116;GO:0005117 GO:0005118;GO:0005119 GO:0005120;GO:0005121 GO:0005122;GO:0005123 GO:0005124;GO:0005177 GO:0005178;GO:0005186	7	3	8	0.167	0.444	0.467	0.700
Inhibitors	12	GO:0004857;GO:0008189 GO:0005074;GO:0005092 GO:0008200;GO:0005517	12	3	9	0.125	0.375	0.571	0.800
DNA or RNA associated proteins	3	GO:0003676;GO:0003735 GO:0004748;GO:0003910	255	21	26	0.070	0.086	0.907	0.924

Table 3B. (*Continued*)

Concept	DIAN node number	Highest level matching GO modes	Present in			Variation		Sensitivity	Selectivity
			DIAN and GO	DIAN only	GO only	Type I	Type II		
		GO:0003911;GO:0004518 GO:0003899;GO:0008534 GO:0008263;GO:0003907 GO:0003905;GO:0003906 GO:0003904 GO:0004844;GO:0003908							
Protein secretion and chaperones	13	GO:0003754;GO:0008565 GO:0006605	11	3	2	0.188	0.125	0.846	0.786
Electron transport proteins	5	GO:0005489	0	7	6	0.538	0.462	0.000	0.000
Other tranport proteins	6	GO:0005215	62	17	19	0.173	0.194	0.765	0.785
Structural proteins	7	GO:0005198	31	23	40	0.245	0.426	0.437	0.574
Receptors	8	GO:0004872	67	43	15	0.344	0.120	0.817	0.609
Cytokines and growth factors	9	GO:0008083;GO:0005125 GO:0008009	35	19	10	0.297	0.156	0.778	0.648
Total DIAN and GO: 480 Total DIAN only: 139 Total GO only: 135 Total: 754 Average type I: 0.184 Average type II: 0.179 Sensitivity: 0.780 Selectivity: 0.775									

A group of well-characterized, nonredundant mouse sequences were assigned to the Protein Function ontology by the DIAN domain-based mapping algorithm. These assignments were compared to assignments made to the GO Process and Function ontology by the MGI.

investigation of the different evaluation methodologies described here. These are summarized in Table 2. For example, 50% of the *Drosophila* genes were classified against the GO Molecular and Biological Function ontologies by the *Drosophila* community, yet no analysis of the errors associated with this work was presented (Ashburner et al. 2000). This is due to the inherent difficulty of assessing error rates associated with ontological classification, such that none of these genome annotations and their associated evidence codes can be statistically evaluated with confidence levels. Here we provide the first attempt to analyze the error rates associated with ontological classification.

Because of the lack of a collection of comprehensive, robust assignments that can be used as a standard of comparison, it is inherently impossible to achieve a completely robust assessment of any assignment methodology. Consequently, none of the approaches described here were entirely satisfactory because of these fundamental limitations. Problems range from multiple types of biases in testing sets, to the partiality of the field's understanding of the function of the proteins in the test sets. Therefore, in many cases the DIAN algorithms were found to be making plausible assignments that cannot be verified with the present data. Additional problems include variability in the comprehensiveness of assignments made to a given protein, as well as variability in the comprehensiveness of assignments of various proteins to ontologies, that is, differences in the coverage between assignment data sets produced by different methods. For example, in the experiment depicted in Table 4, 40% of assignments generated by DIAN (representing 216 assignments) were originally found to be absent in MGD. These were initially considered to be erroneously introduced by the DIAN algorithm, and were

Table 4. Requirement for Manual Validation of Comparative Results

Concept	DIAN/Role node number	Present in DIAN and GO	DIAN only	GO only	Reported type I variation	Effective number of type I assignments	Effective rate of type I variation
Cytoskeletal	3.1	30	19	16	0.29	4	0.06
Nucleotide	6.5	6	25	11	0.60	3	0.05
Sugar/glycolysis	6.7	0	35	8	0.81	3	0.07
RNA polymerases	5.1.1	0	4	0	1.00	4	0.00
RNA processing	5.1.2	4	9	10	0.39	1	0.04
Transcription factors	5.1.3	142	124	98	0.34	0	0.00**

Type I variation here refers to those assignments made by DIAN but not in the reference ontology implementation system (GO system). Manual validation results show that Type I variation (DIAN-only assignments) cannot simple be treated as Type I error in a strict statistical sense.

therefore classified as type I variations. However, on manual review, most of these assignments were found to be correct, such that the number of true type I variations was ultimately reduced to 2.5%. Thus the Type I and II variations in our evaluation scheme cannot be interpreted simply as Type I or II errors in a strict statistical sense. The missing assignments presumably reflect limitations in the keyword-recognition algorithm used in most of the assignments currently provided by the Mouse Genome Database (outlined in Fig. 5A). As an illustration, MGD assignments for entry #104661, which codes for RAR-related orphan receptor α, are depicted in Figure 5B. This gene, a member of the nuclear hormone receptor superfamily involved in thyroid hormone signaling pathway, was assigned to GO categories by MGD on the basis of electronic annotation using a keyword-scanning algorithm (GO evidence code IEA). This algorithm correctly identified the protein function as "DNA binding" and the role of the gene as "transcriptional regulation", but failed to also indicate its receptor function, which is involved in cell signaling (Fig. 5B).

Despite the fact that a more systematic evaluation of assignment algorithms is not feasible because of these deficiencies, results from the evaluation approaches applied here indicate that DIAN returns generally correct assignments of proteins to its various ontologies. Deficiencies in DIAN's assignment algorithms were most manifest in its favoring of underprediction (type II variation). Our manual curation and validation indicate that this error type is far more common than overprediction. This reflects the conservativeness of the selection of protein domains as bona fide annotation signatures for a given node, as well as the limited coverage of the protein universe by domains presently available in the PFAM database, on which the current version of the algorithm is based. In contrast, overprediction is much less frequent and relates to domains that are not completely specific to a given concept and thus return spurious assignments. Other problems include limitations in the resolution of the algorithm, such that DIAN may be unable to correctly assign sequences to very specific nodes such as leafs in the Enzyme hierarchy.

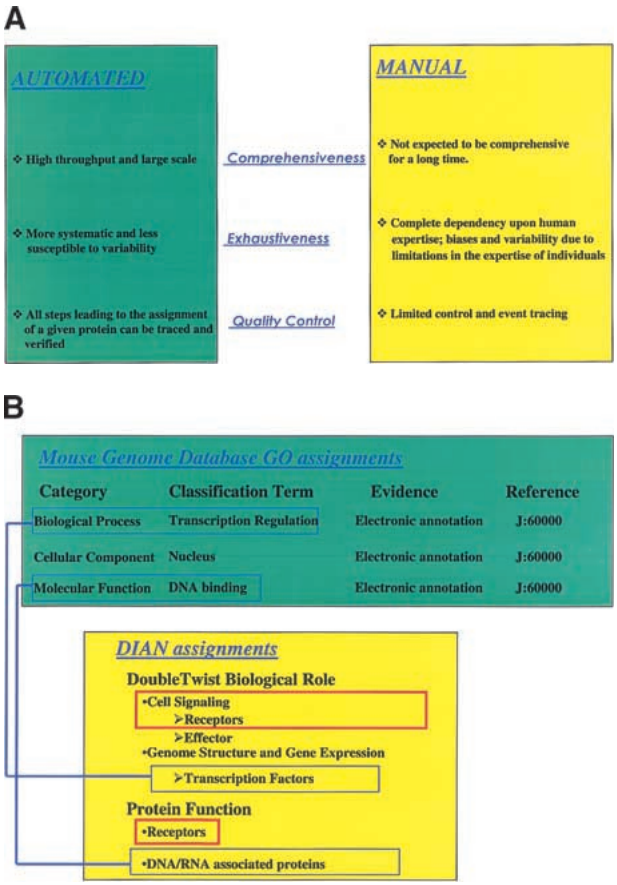

Figure 5 Comparison of assignment methods. (*A*) Comparison of automated and manual assignment methods. The properties of automated assignment methods such as DIAN are compared with those of manually generated assignments. (*B*) Comparison of DIAN and MGI assignments. Results from a simple keyword-based method are illustrated here in assignments made by the algorithm used by Mouse Genome Informatics Database, as compared with DIAN assignments. Note that the "DNA binding" cellular role is vague, as the correct function for this gene should be "transcription factor."

DISCUSSION

Considerations Related to the Assignment of Protein Domains to Biological Ontologies

Because protein domains often involve many-to-many relationships with respect to biochemical function, that is, a given domain may be associated with multiple biochemical functions, the importance of curating these associations to ensure specificity is essential to reduce incorrect assignments. This is most manifest in cases where a simple linkage is made between a protein domain to a biological ontology, such as in the PRINTS and PROSITE databases. Therefore, it is necessary to review the specificity of an assignment in the context of all other assignments this domain may have to other nodes. Furthermore, an evaluation of a nodal annotation signature with respect to the protein universe, here currently approximated by the SWISS-PROT database, is required to be statistically rigorous. For such a review to be robust, it becomes necessary to first associate all known domains to all protein functions described by an ontology, followed by estimating the significance of these associations to ensure that they are informative and not due to, for example, a requirement for a structural role unconnected to the protein function under consideration. This is because only a fraction of domains are truly diagnostic for a given protein function, and although careful manual review can help strengthen the quality of these associations, we believe that only when a global view of associations is available can domains with a low specificity to different functions be eliminated and meaningful assignments be made. Because of the magnitude of the work, generating such a global view can only be achieved via a combination of automation followed by manual curation.

In the case of DIAN, this was accomplished by deriving manually a knowledge base composed of the assignments of all SWISS-PROT proteins to the various ontologies used by DIAN. This was to serve as the first step in defining domains that are meaningfully associated with protein function. This knowledge base was then used to perform exhaustive verifications of the significance of these associations by deriving a heuristic decision rule by which to accept or reject the association of individual domains to ontological nodes. For each candidate protein domain for the annotation signature of the ontology node, the annotations of all SWISS-PROT sequences containing this particular protein domain were analyzed against the SWISS-PROT sequences previously associated with

this node by the DIAN knowledge base. The significant overlap between these pools of SWISS-PROT records and PFAM domains ensures that a particular protein domain can be used as a nodal signature. This information is enabled in a heuristic rule that further requires that a majority of at least four of five SWISS-PROT proteins used in the knowledge base be nonfragmentary, and that annotations associated with these sequences be derived from published laboratory results.

Evaluation of Keyword-Based Versus Domain-Based Ontology Nodal Assignment Methods

As described earlier, DIAN combines two algorithms for the automated assignment of proteins to ontologies that rely on an underlying knowledge base assembled using manual curation, along with heuristic rules. By comparison, other assignment efforts, such as those made by MGD in the context of the GO consortium, currently rely primarily on a simple process of scanning source records for keywords to GO terms. Full manual assignment of records is intended to follow this initial phase. However, such human curation poses several significant limitations, among which is the prohibitive expense of genome-scale assignment. For this reason, over 84% of the 14,801 assignments presently available in MGD were generated by using keyword-based association, with the remaining assignments being produced manually. Because automated assignments methods can be expected to remain a key technology due to their high-throughput capability, development of algorithms that go beyond the limitations of simple keyword-based assignment is imperative, as most genes will not receive the kind of textual descriptions that lend themselves to this approach. Therefore, the domain-based approach of DIAN provides a distinct additional approach beyond keyword scanning, and permits high-throughput assignment independently of the presence of prior textual annotations, while retaining reasonable accuracy. Lastly, because of the frequent difficulty of confirming whether a given assignment is incorrect, such reviews should perhaps be limited to providing a general confidence value on the mappings made by automated methods. As was done here, selective manual reviews of individual assignments based on the comparison of different algorithmic implementations can also be used to uncover possible errors and defects in their respective mapping methodologies. Worthy of mention here is DIAN's validation module, which integrates manual reviews to compensate for deficiencies in the various automated validation methods.

In summary, DIAN is a high-throughput annotation algorithm that uses biological ontologies to segment the spaces of protein function, biological role, and structure. When applied to data generated from genome sequencing projects, DIAN is an effective algorithm for the conceptual annotation of genome-scale in a timely and scientifically accurate manner. It is also an effective data mining algorithm, applicable to the identification of novel correlations that can only be made at the conceptual level.

METHODS

Ontologies

DIAN currently supports five ontologies: the PROSITE ontology was used for Protein Function (http://www.expasy.ch/prosite/, release/version 16.30); Cellular Role is enabled by the EGAD ontology from TIGR (http://www.tigr.org/docs/tigr-scripts/egad_scripts/role_report.spl, release/version N/A), which was originally derived from Monica Riley's *E. coli* protein ontology; the Enzyme classification is from IUBMB (International Union of Biochemistry and Molecular Biology (http://www.chem.qmw.ac.uk/iubmb/enzyme/, release/version Enzyme Nomenclature 1992 and all of its supplements); SCOP is from the Medical Research Council (MRC) of the United Kingdom (http://scop.mrc-lmb.cam.ac.uk/scop/index.html, release/version 1.53); the DoubleTwist Biological Role was derived internally (release/version 1.00). These ontologies can be viewed as taxonomies of IS-A links, in which a node situated at level 1 (Fig. 1) indicates a node expressing a more general concept than that of a node at level 2, whereas a node situated at level 3 indicates a more specialized node than the one at level 2.

Component Databases Supported by DIAN

The component databases supported by DIAN are the GenBank primate division (GB Release 121); UniGene (Build #129); SWISS-PROT (Release 39); PDB (Release as of 1/1/2001); and GenPept (Release as of 12/27/2000).

Construction of the DIAN Knowledge Base

Two databases were constructed as the foundation of the knowledge base associated with ontological nodes: a controlled vocabulary and regular expression database, and a protein domain signature database. For classification of protein structures, the PFAM motifs within the SCOP domain sequences compiled by the SCOP consortium (Brenner et al. 1998) were used as source material for the nodal signatures of the structural ontology. The controlled vocabulary database was populated during the construction of the SWISS-PROT-Ontology mapping table. A computer-aided human curation process was performed by a group of domain specialists whereby SWISS-PROT sequences were manually assigned to the supported ontologies. Node-specific vocabulary and regular expressions were derived and later used to control the association of source-record annotations to a given node of the supported ontologies. In this way, vocabulary data sets for each relevant node were created and manually curated with subsequent releases of SWISS-PROT. Using the manually curated association between SWISS-PROT sequences and ontological nodes, sequences in this database were processed to identify PFAM protein domains using Paracel's GeneMatcher system. Through the SWISS-PROT-Ontology table, annotations made with respect to PFAM domains in SWISS-PROT source records were used to verify the accuracy of the association of PFAM domains to an ontology node before assigning a domain to a node. Specifically, because PFAM domain and ontology node each have a satellite pool of SWISS-PROT records, the extent of the overlap between these pools of records is used to confirm the correctness of the assignment of this PFAM domain to a particular ontology node. This was done in a many-to-many manner, such that a domain can be assigned to more than one node, and a given node can have more than one domain associated with it.

DIAN Algorithm Implementation

The underlying DIAN knowledge base was implemented using the Oracle 7.3 relational database management system (Oracle). For Hidden Markov Model searching, the GeneMatcher system was selected for its ability to perform high-throughput protein domain profiling using the PFAM database. User queries of the DIAN data set are performed using the PLS index-based search engine (http://www.pls.com) from American Online. Most of the DIAN pipeline was implemented using the Perl (v.5.0) language. Benchmarks of chromosome 22 were obtained as follows: chromosome 22 was first fragmented into overlapping fragments of 200,000 bp. GENSCAN (Burge and Karlin 1997) was used to generate a da-

tabase of predicted gene sequences. This collection of gene predictions was then processed by the DIAN pipeline for annotation analysis. In this case, rather than using GeneMatcher, PFAM domain profiling was done by farming the predicted gene translations to four workstations running the HMMER software package to show that DIAN can be applied easily as a component of a large-scale annotation system for genome-scale sequencing projects using a conventional computing architecture. The coverage by DIAN of chromosome 22 was thus based on this database of predicted gene sequences. Only the domain-based assignment algorithm was used in this case.

DIAN Algorithm Evaluation

Three approaches were applied in evaluating the assignment accuracy of DIAN: manual verification, comparisons between assignments to different ontologies, and ortholog-based validation. In the first approach, manual verification of assignments was made to selected proteins. A group of domain experts was given the task of reviewing annotation assignments of biological sequences made by the DIAN pipeline within their domain of expertise. Several dozen proteins of varied types were evaluated in this manner. In the second approach, a test set was constructed for the comparative evaluation of assignments. Nodes from the GO:Process or GO:Function ontologies that are conceptually equivalent to nodes of the DIAN Protein Function or Cellular Role ontologies were identified (Table 3, Fig. 4A,B; see Fig. 4 for explanation). Mouse genes assigned by MGD (http://www.informatics.jax.org/; Baker et al. 1999) to these GO nodes (or their child nodes) were then retrieved. The protein sequences for these genes were obtained from RefSeq via shared HUGO gene names (http://www.gene.ucl.ac.uk/nomenclature/). An all-versus-all Smith-Waterman sequence similarity search (Smith and Waterman 1981) was further performed to eliminate sequence redundancy within these mouse sequences. Only sequences with &lt:40% overall similarity were retained as the testing set, composed of 857 proteins. These sequences were then assigned to DIAN ontologies by the DIAN algorithm for comparison against their original assignments in GO ontologies (Table 3A, 3B). Sequences with unbalanced assignments between GO and DIAN ontologies were examined manually to assess the source of the imbalance: the presence of a missing assignment of a bona fide property listed in GO, or a missing or incorrect assignment of a bona fide property in DIAN. In the last approach, assignments made to orthologous sequences were compared. A test set of orthologous proteins was assembled, composed of a random set of 37 pairs of orthologous Refseq protein records for mouse and human. Orthology was assumed when genes shared the same HUGO gene name. Sequences from the test set were processed by the DIAN pipeline, and resulting assignments were compared between proteins, with the expectation that identical assignments would be generated. Sequences with unbalanced assignments were examined manually to assess the source of the imbalance, such as the presence of a species-specific function or from a possibly erroneous assignment made by the DIAN algorithm.

ACKNOWLEDGMENTS

A patent application for the DIAN algorithm has been filed with the U.S. Patent and Trademark Office. The authors are grateful to Drs. Doug Brutlag (Stanford University), Peter Karp (SRI International), and Andrew Karsaskis (DoubleTwist, Inc.) for valuable discussions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25:** 25–29.

Attwood, T.K. 2000. The Babel of bioinformatics. *Science* **290:** 471–473.

Attwood, T.K., Avison, H., Beck, M.E., Bewley, M., Bleasby, A.J., Brewster, F., Cooper, P., Degtyarenko, K., Geddes, A.J., Flower, D.R., et al. 1997. The PRINTS database of protein fingerprints: A novel information resource for computational molecular biology. *J. Chem. Inf. Comput. Sci.* **37:** 417–424.

Bairoch, A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res.* (Suppl.) **19:** 2241–2245.

Bairoch, A. and Apweiler, R. 2000. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28:** 45–48.

Baker, P.G., Goble, C.A., Bechhofer, S., Paton, N.W., Stevens, R., and Brass, A. 1999. An ontology for bioinformatics applications. *Bioinformatics* **15:** 510–520.

Ben-Yaacov, S., Le Borgne, R., Abramson, I., Schweisguth, F., and Schejter, E.D. 2001. Wasp, the Drosophila Wiskott-Aldrich syndrome gene homologue, is required for cell fate decisions mediated by Notch signaling. *J. Cell Biol.* **152:** 1–14.

Birney, E. and Durbin, R. 2000. Using GeneWise in the Drosophila annotation experiment. *Genome Res.* **10:** 547–548.

Blake, J.A., Eppig, J.T., Richardson, J.E., and Davisson, M.T. 2000. The Mouse Genome Database (MGD): Expanding genetic and genomic resources for the laboratory mouse. The Mouse Genome Database Group. *Nucleic Acids Res.* **28:** 108–111.

Bowie, J.U., Luthy, R., and Eisenberg, D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253:** 164–170.

Brenner, S.E., Chothia, C., and Hubbard, T.J. 1998. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl. Acad. Sci.* **95:** 6073–6078.

Burge, C. and Karlin, S. 1997. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.* **268:** 78–94.

Chen, R.O., Felciano, R., and Altman, R.B. 1997. RIBOWEB: Linking structural computations to a knowledge base of published experimental data. *Ismb* **5:** 84–87.

Chervitz, S.A., Hester, E.T., Ball, C.A., Dolinski, K., Dwight, S.S., Harris, M.A., Juvik, G., Malekian, A., Roberts, S., Roe, et al. 1999. Using the Saccharomyces Genome Database (SGD) for analysis of protein similarities and structure. *Nucleic Acids Res.* **27:** 74–78.

Commission on Biochemical Nomenclature, and International Union of Biochemistry. Standing Committee on Enzymes. 1973. *Enzyme nomenclature; recommendations (1972) of the Commission on Biochemical Nomenclature on the nomenclature and classification of enzymes together with their units and the symbols of enzyme kinetics*. Elsevier Scientific, New York.

Corpet, F., Gouzy, J., and Kahn, D. 1998. The ProDom database of protein domain families. *Nucleic Acids Res.* **26:** 323–326.

Derry, J.M., Ochs, H.D., and Francke, U. 1994. Isolation of a novel gene mutated in Wiskott-Aldrich syndrome. *Cell* **78:** 635–644.

Gracy, J. and Argos, P. 1998. Automated protein sequence database classification. II. Delineation of domain boundaries from sequence similarities. *Bioinformatics* **14:** 174–187.

Gusfield, D. 1997. *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge University Press, Cambridge.

Henikoff, S., Henikoff, J.G., and Pietrokovski, S. 1999. Blocks+: A non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics* **15:** 471–479.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

International Union of Biochemistry. Standing Committee on Enzymes. 1965. *Enzyme nomenclature; recommendations, 1964, of the International Union of Biochemistry on the nomenclature and classification of enzymes, together with their units and the symbols of enzyme kinetics*. Elsevier, New York.

International Union of Biochemistry. Nomenclature Committee and Commission on Biochemical Nomenclature. 1979. *Enzyme nomenclature, 1978: Recommendations of the Nomenclature Committee of the International Union of Biochemistry of the nomenclature and classification of enzymes*. Academic Press, New York.

International Union of Biochemistry. Nomenclature Committee, International Union of Biochemistry, and Commission on Biochemical Nomenclature. 1979. *Enzyme nomenclature, 1978: Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the nomenclature and classification of enzymes*. Academic Press, New York.

International Union of Biochemistry. Nomenclature Committee, Webb, E.C., and International Union of Biochemistry. 1984. *Enzyme nomenclature 1984: Recommendations of the Nomenclature Committee of the International Union of Biochemistry on the nomenclature and classification of enzyme-catalysed reactions*. Academic Press, Orlando, FL.

International Union of Biochemistry and Molecular Biology. Nomenclature Committee and Webb, E.C. 1992. *Enzyme nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes*. Academic Press, San Diego.

Karp, P.D. 2000. An ontology for biological function based on molecular interactions. *Bioinformatics* **16:** 269–285.

Lewis, S., Ashburner, M., and Reese, M.G. 2000. Annotating eukaryote genomes. *Curr. Opin. Struct. Biol.* **10:** 349–354.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247:** 536–540.

Nevill-Manning, C.G., Wu, T.D., and Brutlag, D.L. 1998. Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci.* **95:** 5865–5871.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. 1999. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **27:** 29–34.

Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., and Thornton, J.M. 1997. CATH—A hierarchic classification of protein domain structures. *Structure* **5:** 1093–1108.

Rawlings, S.L., Crooks, G.M., Bockstoce, D., Barsky, L.W., Parkman, R., and Weinberg, K.I. 1999. Spontaneous apoptosis in lymphocytes from patients with Wiskott-Aldrich syndrome: Correlation of accelerated cell death and attenuated bcl-2 expression. *Blood* **94:** 3872–3882.

Rengan, R., Ochs, H.D., Sweet, L.I., Keil, M.L., Gunning, W.T., Lachant, N.A., Boxer, L.A., and Omann, G.M. 2000. Actin cytoskeletal function is spared, but apoptosis is increased, in WAS patient hematopoietic cells. *Blood* **95:** 1283–1292.

Riley, M. 1993. Functions of the gene products of *Escherichia coli*. *Microbiol. Rev.* **57:** 862–952.

Smith, T.F. and Waterman, M.S. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* **147:** 195–197.

Sonnhammer, E.L., Eddy, S.R., Birney, E., Bateman, A., and Durbin, R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26:** 320–322.

Wheeler, D.L., Church, D.M., Lash, A.E., Leipe, D.D., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Tatusova, T.A., Wagner, L., et al. 2001. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **29:** 11–16.

Wu, T.D., Nevill-Manning, C.G., and Brutlag, D.L. 2000. Fast probabilistic analysis of sequence function using scoring matrices. *Bioinformatics* **16:** 233–244.

Received February 7, 2001; accepted in revised form August 14, 2001.